15th International Roundtable on Business Survey Frames
Washington, D.C. – October 22 – 26, 2001

Session No 6
Paper No 3
Giuseppe Garofalo, ISTAT, Italy

# To Exploit Administrative Sources: A Framework of Concepts

## *1. Introduction*

In recent years, the issue of using administrative sources more extensively for statistical purposes has moved noticeably higher in many country. The reasons for this quickening of interest can be summarised as follows:

- National Statistical Institute are trying to meet more exacting demands, by governments, international bodies and researchers, for business statistics, in particular as regards of microeconomic analysis for small areas, specific sectors of activity and for particular aggregates. At the same time they are under pressures to reduction their own collection costs and reporting burden.

- Rising proportions of GDP and employment are being contributed by small enterprises, for which sample surveys are often not easy to conduct. Always their efficiency is less than the reliable estimation of variables taken from other sources.

- Recent advances in information technology have made the large arrays of data characteristic of administrative sources a great deal easier to handle and have opened up new possibilities for linking different statistical and administrative databases.

For these reasons recent National and Community legislation, in statistical aspects, are focused their aim in the regulation of access of administrative databases for statistical purposes.

For general aspects the primary reference is the "CRCS regulation[1]", so-called "Statistical Low". Chapter V prescribes in article 16 that: in order to reduce the burden on respondents… the National authority and the Community authority shall have access to administrative data sources, each in the of activity of their own public administration". For specific applications authorisations, to collect data contained in the administrative or legal files, are regulated in the Council Regulations concerning the creation and maintenance business registers[2], the structural business statistics[3], the short term business statistics[4].

As regards the Italian legislation, the 1989 degree requires all public bodies to provide information needed for national statistical program. This degree is not enough in practice to secure access to particular administrative sources. The 1996 law which authorised the current Intermediate Census of Industry and Services obliges public bodies to give statisticians access specifically to administrative registers, archives and micro-data.

Having legal bases is necessary but not enough. The Canadian statistician G. J. Brackstone wrote: "it may be less important to have a watertight definition than to have an understanding of the features that distinguish administrative data from data from statistical sources in the context

[1] Council Regulation N° 322/97 of 17 February 1997 on Community Statistics.
[2] Council Regulation (EEC) N° 2186/93 of 22 July 1993 on Community co-ordination in drawing up business registers for statistical purposes.
[3] Council Regulation (EC) N°58/97 of 20 December 1996
[4] Council Regulation (EC) N°1165/98 of 19 May 1998.

of statistical use"[5]. He suggested four features: the agent and the unit are different; the data are collected for a non-statistical purposes; the coverage of the target population; control of the method by which the administrative data are collected and processed rests with the administrative agency. As Brackstore commented, each of these features affects the character of administrative data and has implications for the use of administrative data for statistical purposes.

In this paper a "general" conceptual approach is presented for the use of administrative sources for statistical purposes. The problems connected to the quality of administrative data and basic conceptual criteria to solve the problems of the inconsistency of administrative information (as compared to the statistical one) are evidenced. Finally limits and connected problems to be solved with particular reference to the use of integrated administrative data with the aim to set-up and update a Business Statistical Register (BR), are individuated.

**2. The use of administrative source: inconsistency of the data.**

Enterprises systematically produce administrative and legal acts during their lives: they pay taxes, stipulate telephone and electric contracts, insure the employees against accidents on work, etc. All these are administrative acts but in reality each of them hides information that is useful to locate and explain in statistical point of view.

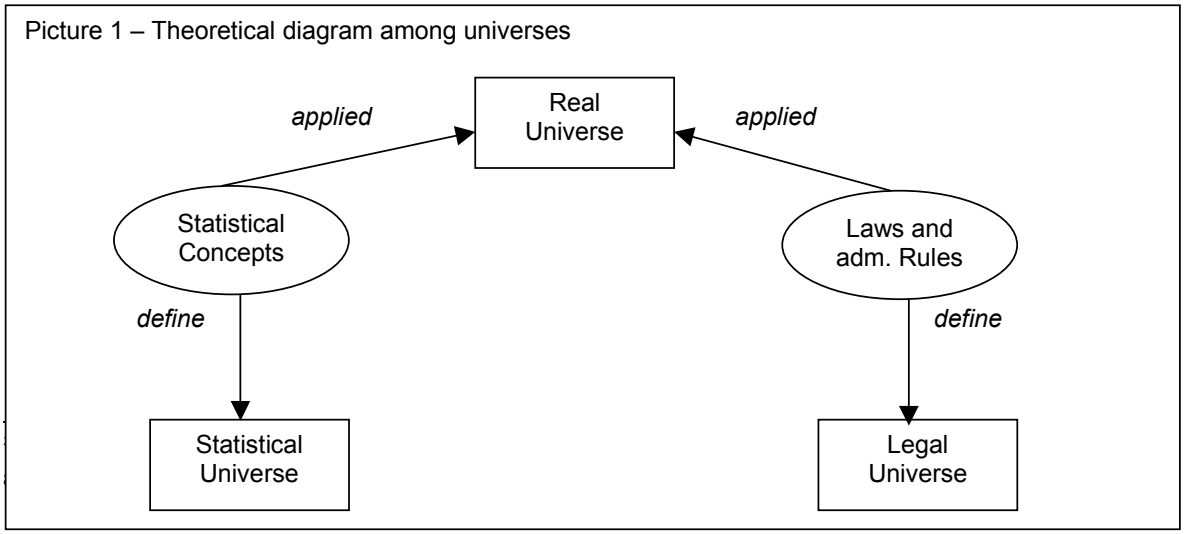
Every administrative body has its own function to collect data and manage the corresponding records, under specific legislation and rules which govern relations between various individuals and between them and the public administration. Thus, each source makes use of definitions, classifications and rules on entry and cessation that are peculiar to itself and depend on the functions of the authority concerned. The administrative body defines, classifies, collects and records information on economic persons and their characteristics that, in the strict sense of word, do not have statistical validity. In other words using administrative data causes statisticians a problem (not of easy solution): the inconsistency of data.

From a general point of view the objective of statistics is to analyse the real world's phenomena using own definitions and concepts. On the other hand the real world's phenomena are ruled by rules and laws governing relationships between persons (physical and juridical) and between those and the public administration. The "administrative laws and rules" and "statistical concepts" determine two different images of the real universe, which we can define respectively:

- *Legal universe*, composed by legal units
- *Statistical universe*, composed by statistical units.

Picture 1 schematises the relations between the different defined universes. From a conceptual point of view no direct relation exists between the legal universe and the statistical one: they are two different "logical views" of the same object: the real universe.

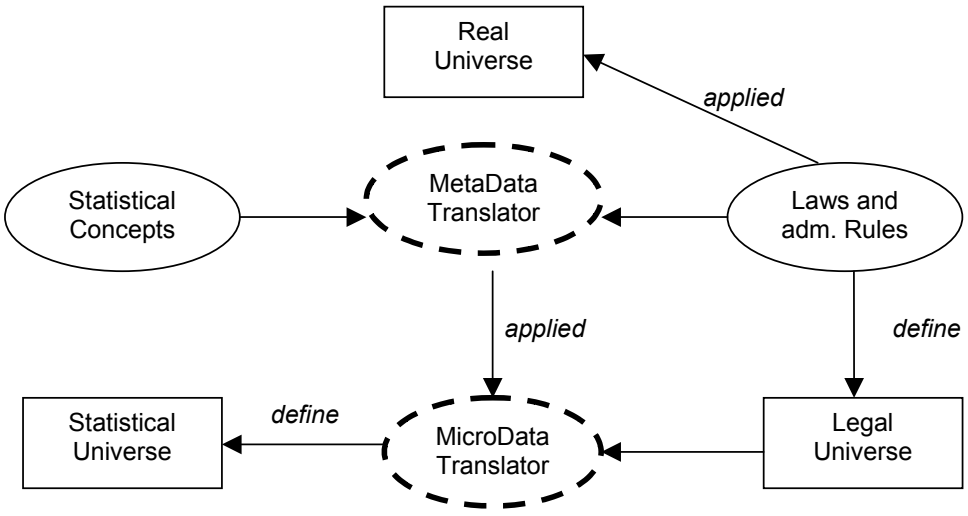
Picture 1 – Theoretical diagram among universes



In this sense to use of administrative files (where legal units are registered) for statistical purposes, means to use an image of the real universe "mediated" by the administrative laws and rules of the legal universe.

So the main problem that arises through the use of administrative sources for statistical purposes is to identify the correspondences (MetaData Translator) between the statistical concepts and the administrative rules and laws through which those sources observe the universe, or population, of reference. It is therefore necessary to handle in some way the administrative sources in order to align them to the statistical concept and definitions. This is possible if, at one side, we have a deep knowledge of the sources to be used and, on the other side, suitable statistical methodologies (MicroData Translator) are available (Picture 2).

Picture 2 – Diagram of the relationship between legal and statistical universe




It is possible to synthesize the logical process in order to use information deduced by administrative sources, according to the following three conceptual steps:

1. *definition of the reference conceptual frame*: statistical definitions and classifications;
2. *knowledge of the observed universe* (administrative files) refering to coverage, definition of the units and characters, classification used, time and modalities of updating;
3. *identification of the MetaData Translator* to convert administrative data in statistical ones.
4. *develop of the suitable statistical methodologies* in order to treat administrative information

### *3. Quality problems in the use of administrative sources*

#### *3.1 The general conceptual model*

The use, with statistical goals, of the information contained in an administrative basis of data (for example the fiscal register, used in the majority of European Countries as a priority source in the realisation and update of the BR) can cause serious problems:

- The source used may have an informative covering grade, in terms of units and variables, different from the one of the statistical reference. The covering grade is determined not only by the definition of the persons obliged to registration in the specific administrative register, but also by the "conventions" used for the enrolment/cancellation, and by the delays with which they are registered.

- The prescribed and management legislative variations, which are determined in the course of time, may modify in a non secondary way the informative content of the administrative source so to determine spurious structural *breaks* (in the deriving statistical product) in terms of false demographic fluxes.

- The rules and the administrative practices evolve in time influenced by political intervention on the incentive processes of economy and of accomplishment of the fiscal policy. The economical operators use legal concepts and modify them, in practice, on the basis of their own interests. It is not only a question of "informal, underground or illegal economy" but, with the aim to partially elude the public revenue or to use facilitated subsidies and funding, behaviours, taking to "declarations" not corresponding - partially or totally - to the operator's "economic reality", are developed.

The elements previously mentioned show how the use of administrative data for statistical purposes, impose the necessity to solve the usual problems of a statistical survey - accuracy, completeness, timeliness - by a new conceptual and methodological approach.

Within a survey (or a system) for the collection of statistical data quality is a problem evaluated ex-ante as well as it is strongly linked to the process of microdata collection and macrodata production. When we want to use data stored in non-statistical (administrative) databases, for which statisticians do not have any control of the production process, the problem of consistency is set in a different context and it is resolvable only ex-post. Only an appropriate use of statistical methodologies can assure consistent results.

Referring to a defined statistical universe, the typologies of errors generated in the use of a specific administrative source for statistical purposes, can be summarised as follows:

| | |
|---|---|
| *E1 – error of under-recording* | a) missing-recording of legal persons due to delays of iscription, etc…<br>b) undrecording of legal person not obliged to the registration.<br>c) underecording of legal unit due to fiscal elusion and evasion |
| *E2 – error of over-recording* | a) registration of not active legal person due to duplications, delays of cessation.<br>b) registration of legal persons without any feature of enterprise<br>c) legal units only formally different (duplication for fiscal and administrative reasons) |
| *E3 – error assignment of characters* | a) wrong recording due to delays in variations acquired or to errors in declarations, in recording, in checking.<br>b) wrong recording due to different definitions and classifications<br>c) voluntary mistakes of classification by the legal persons. |
| *E4 – missing assignment of char*. | a) partial or total missing data |

Let $x_i$ the real value of the i-ma unit related to the attribute X (eg. economic activity code) and $x_{ij}$ is the values recorded in the admininistrative source (observed value). The relation between the available values and the real ones can be described as follows:

$$x_{ij} = g(x_i, e_j, \varepsilon_{ij}) \quad (1)$$

The observed value ($x_{ij}$) is a function of the real value ($x_i$) and the erratic component described by the three different typologies of errors:

1. $\varepsilon_{ij}$ is the random error (errors type “a”).
2. $e_j$ is the structural error due to the characteristics of the source (errors type “b” and “c”)

*3.2 The adjustment of the systematic error*

In presence of a complete knowledge of the administrative source it is relatively easy to individuate and adjust the systematic mistake connected to the characteristics typical of the source used. The adopted method is the individuation of rules which standardise (or harmonise or normalise) the units and the characters of input source in statistical units and variables. So it is possible define the *standardisation function* as the following application:

$$f_s: X_j \Rightarrow X$$

which changes the values $x_{ij} \in X_j$ in values $x_i \in X$. In other words, a standardisation function converts administrative concepts and classifications into statistical ones.

This rule, generally deterministic, can be divided into three types:

- *coding rules*: which convert coding (e.g. economic activity, legal form, and location) into statistical classifications (Nace, Nuts, etc. );
- *link rules*: by which the different records corresponding to legal or administrative units in one source can be combined to define one statistical unit (enterprise or local unit);
- *conversion rules*: to obtain statistical variable from administrative characters

But the situation is more complicated. The systematic errors in each sources, produced by administrative-juridical functions, are the result of two separate elements: the first tied to laws, classifications, proper definitions of the source (error type “b”), can be located and easily standardised; the latter, tied to elusion and evasion phenomena (error type “c”), can’t be located and is misted inside the random error. In such way the model will change and be complex as follows:

$$x_{ij} = g(x_i, \delta_j, \gamma_j, \varepsilon_{ij}) \quad (2)$$

where $e_j = h(\delta_j, \gamma_j)$ is the result of a know an unknown systematic component. The bias which weight on the definition of the input values of administrative sources is hardly quantified and well known in literature. A clear example is the trend of enterprises in some trade sectors to be classified, for fiscal facilities or credit access reasons, as manufacturing enterprises. It is also well-known the trend to subscript in the social security register, persons not carrying out subordinate employment in order to ensure a social security insurance.

*3.3 The integrated approach to solve the quality problem*

For the above mentioned reason together with the conceptual steps previously located, it is necessary (when the costs are admissible) to develop the further function of "identification of rules for the integration of data coming from more administrative sources".

The integration process is useful when the considered input sources do not assure the completeness in units and in characters’ units, obtaining in such way a reduction of errors of type E1 and E4. Such process must be less useful for the reduction of over recording errors and of wrong character attribution. In fact using more sources can cause an increase of type E2 error; while if each source is really and considerably better of other, further information for imputation of statistical characters would cause troubles. Besides the presence of not checked matching procedures among sources could cause record duplications and therefore an overestimation of units and statistical aggregates.

Formally speaking, within the integration process, the available data have the structure shows in the below table:

| Units | Sources | | | | | Real Value |
|---|---|---|---|---|---|---|
| | 1 | ..... | J | ..... | M | |
| 1 | $x_{11}$ | ..... | $x_{1j}$ | ..... | $x_{1m}$ | $x_1$ |
| 2 | $x_{21}$ | | - | | - | - |
| 3 | - | ..... | $x_{3j}$ | ..... | - | $x_3$ |
| ..... | ..... | ..... | ..... | ..... | ..... | ..... |
| | . | | | | . | |
| I | $x_{i1}$ | ..... | $x_{ij}$ | ..... | $x_{im}$ | $x_i$ |
| ..... | ..... | ..... | ..... | ..... | ..... | ..... |
| | . | | | | . | |
| N | $x_{n1}$ | ..... | $x_{nj}$ | ..... | $x_{nm}$ | $x_n$ |

If the sources are independent a similar information structure might suggest the use of statistics based on linear functions (i.e.: the average) to estimate the value $x_i$ . But both the characteristics of the variables to be estimated in a statistical archive (the majority of which have a qualitative character: economic activity, juridical status, activity status) and the complex composition of the mistake component (2) suggest a different approach so to avoid distorted estimates of $x_i$ .

The statistical methodologies adopted in the imputation of characters must be based on the concept “choice among alternative values” and not on “combining the available values”, when there are more values of an attribute for the same unit. The problem is to locate such value $x_{ij}$ which has a minimum value of the unknown error $\eta_{ij} = f(\gamma_i, \varepsilon_{ij})$. The estimation $x_i$ of the real value taken from variable x in the fellow i has been carried out according to the following rule:

$$x_i = x_{ij} \quad : \eta_{ij} < \eta_{ik} \ \forall\, j \neq k \quad \text{e con} \quad k = 1,....,m.$$

In this regard can be adopted statistical methodologies using, in alternative or in a combined system, hypothesis on:

- the real distribution of the character in the population based on the non mis-matching values in the different used sources;

- the quality of the sources, based on previous information when there is a “reliable witness” (for instance data of statistical surveys) that is without systematic errors, or on “ad hoc” indicator for instance the quality of a single source could be quantified by the mean percentage of the errors found by comparing with all the others sources;

- the distribution of the instrumental variables: for instance is possible to exploit information about the enterprise dimension using the turnover or the electrical consumption.

*3.4. The integration of different sources as a necessity for the creation of a BR.*

Independently from the problems connected to the quality, discussed before, the integration of different informative sources is an unavoidable necessity if, to reduce costs and *burden*, statistical surveys cease to be the primary source of information collection. The tendency, by now generalised in the National Statistical Institutes, is to look for information where they are available (administrative files, private or public data banks, informative systems of the same recording units) and to combine them together, using statistical surveys to complete or control the collected information.

Each administrative source often provides partial information, both with regard to the units and the units' characteristics, so it is rarely sufficient to answer to informative needs. The classical theoretic model used, which connected a unique source (the survey) to an informative need is not real any more. The model is transformed in "*an informative need – different sources*"[6] which implies the individuation of new and specific methodologies for the treatment of the information.

A structure of statistical information, determined by an heterogeneous set of sources can be called "*hybrid data collection*" (HDC)[7].

The set-up and the update of a BR takes very often to the characteristics of a HDC. To have a wider accuracy of information, the data collection of the biggest enterprises is obtained using statistical surveys. For smaller units, characteristics are identified using different sources, each of them providing only some of the information needed for the archive. The final result is a complex process determined by:

- *a vertical integration*: the archive units are collected with different techniques (for example big enterprises from statistical surveys, small ones from administrative sources),

- *a horizontal integration*: the units' characteristics are acquired by different sources (for example: the economical turnover from the VAT register and the employees from the register of social security).

In such a situation the risk run is the "internal inconsistency" of the statistical archive and the primary necessity remains the development of a system of concepts and methodologies, described in the preceding paragraphs, the "hybrid" collection of information again to an "internal homogeneity".

***5. The physical data integration***

The methodological and conceptual aspects for the integration of the different administrative database find their premise in the capacity of carrying out a physical connection among records coming from different sources and referring to the same unit.

The record linkage is a complex process of linkage of data relative to the same entity in different files (external linkage) or inside the same file (internal linkage) . In order to practise the linkage automatically, *identification variables* are necessary: identification code, name, address, etc.

Apparently the linkage through the use of identification code is a simple procedure: if the code identifies univocally a unit and it is contained in the file for all units, the operation is a simple "merge". Instead the concept of the record linkage based on not unique attributes (names, aggresses, activity codes,..) becomes more complicated. In fact when the tags are represented by variables expressed in words strings, the process of standardization of each ranges used for the comparison becomes crucial; besides, in case such process had successful and taking as example the use of range where it's recorded the enterprise name, we can think of considering the enterprise in several ways, of the words contained in such range. Obviously the merge of two o more records can't simply indicate there was a linkage among the units. The individuation of the linkage is a decisional and complex procedure, which foresees the results processing of the rules application for the comparison of the matching variables and a valuation of the obtained link connected to the reality.

*4.1 The identification code: the Italian fiscal code as example*

The presence of a unique and self-controlled identification code makes obviously easier the right matching.

---

[6] M. Calzaroni, E. Giovannini, A. Sorce (2000): Il Sistema Informativo Statistico sulle Imprese dell'Istat: problemi e potenzialità, XXX riunione scientifica SIS, Firenze.

[7] P. Rivière (2000) :"Hybrid data collection: towards a general tool for collecting information from heterogeneous sources" 14th Roundtable on Business Survey Frames – Auckland

Following to the experience developed in Italy for the setting-up of a business statistical register, it has been helpful the presence of a identification code, the Fiscal Code, recorded with a fair level of coverage (95%) in all of the sources used. The "fiscal code" is the identification code adopted in the relationship between physical and juridical persons and fiscal administration. It is composed by 16 alphanumerical characters in presence of physical persons and of 11 numerical ones for juridical persons.

For the physical persons the alphanumerical which individuates the code is structured as follow:

- three alphabetical characters for the surname (first, second and third consonant)
- three alphabetical characters for the name (first, third and fourth consonant)
- two numerical characters for the birth year,
- one alphabetical character for the birth mouth,
- two numerical characters for the birth day and the sex (for women the birth day is raised up 40),
- four alphabetical characters for the birth municipality code (or foreign state),
- the sixteen bit is a check figure.

For the juridical persons the code is composed by a numerical expression. The first seven figures represent the registration number of the person within the province of the office assigning the code. The next three figures represent the identification code of the province. Last figure is a check figure

*4.2 The use a probabilistic linkage methodology*

When no unique identifiers exist, it is possible to use matching procedure based on the strings comparison of the identify characters as name and address.

In 1969 Ivan Fellegi and Alan Sunter of Statistics Canada published a paper in the *Journal of the American Statistical Association (JASA)* entitled "A theory of record linkage". This paper provided the theoretical base for much of the developed record linkage (RL) systems and the theory contained in it has become the most widely accepted probabilistic linkage method. Fellegi and Sunter began by assuming the existence of two files of computer records. They assumed that both files contained records representing units from the same population. The authors developed a probabilistic approach to associate the records of the two files that represent the same unit. This approach has the property that a minimum number of *record pairs* have to reviewed by humans to achieve given level of false link and false non-link.

It is assumed each record in the first file would be compared to each record in the second file using a set of identifier components, like name, surname, address, telephone number, birth date, etc. (named matching variables). The outcome is a comparison vector whose components are coded in agreement or disagreement. For each pair of records, a total weigh (w) is assigned so that if w is large enough, the pair is defined as link, if the w values is small, the pair is defined as non-link. Fellegi and Sunter suggested a general form for a decision rule that permits to identify thresholds for given levels of errors which divides the set of outcome vectors into three subset: links, non-links, possible links The decision rule can be defined so that the set of possible links as small as possible, that is, the system automatically makes decisions about many record pairs as possible.

The probabilistic linkage methodology has been applied successfully, especially in USA and Canada, in order to integrate demographic, social and health data. Some generalized software packages are developed[8].

The use of these procedures within the matching of enterprises files is still partial and only in the last years realisations have been carried out in the agricultural sector in Canada and in USA.

The major difficulties caused during the development of the probabilistic linkage are originated from the complex structure of identification characters used as matching variables. In

[8] AutoMatch/Autostan, developed by MachWare Technologies Inc. - USA, and GRLS, developed by Statistics Canada

comparison with individuals a business has more complex variables both in their structure and in the meaning. For example, referring the most important used variables:

| | |
|---|---|
| Individual: | Date of birth (unique) |
| Business unit : | Date of inscription (not unique , it depends on the administrative source come from) |
| Individual : | Surname and name (a simple structure) ; |
| Business unit : | Enterprise name (a complex structure, especially for the companies) |
| Individual : | Address (home address unless moving) ; |
| Business unit: | Address declared (legal place); administrative office (where the administrative offices are); establishment (where the activity is carried at local level) |

### *6. Some critical issues in the use of administrative data*

The use of administrative sources for the set-up and the update of a BR presents a series of critical issues. Following we are indicating the most relevant.

1. *The knowledge of the sources used.* We have underlined before how the priority aspect to be faced is the knowledge of the informative content of administrative sources: the legal basis, the persons involved, the definitions adopted, the methods and times to collect information, the treatment of data. A correct knowledge of the acquired information surely facilitates its use and its correct treatment. Very often we are at the presence of rules not always correctly codified, varying quickly in time, because of legislative modifications, which can be different for different territorial areas. For this reason it is necessary:
   - to develop a precise and continuous monitoring activity
   - to intervene on the administrations, providing information, at least for the aspects related to the forms used and the classifications adopted.
   - to develop control procedures on the contents and the quality of the information collected.

   This aspect is very often disregarded because it is considered as secondary and as a useless, expensive and additional burden.
   It is anyway always necessary to define a “protocol” for the acquisition of administrative sources intended as a “minimum set of necessary meta-information” for the correct management of the data. The content s of this protocol must be:
   - General legislative frame: set of the laws, decrees, rules referred to the information acquired.
   - Main function of the administrative archive
   - Persons obliged to enrolment
   - Rules of Enrolment/cancellation
   - Forms used
   - Definition of the registered data
   - Times and procedures for the update of the data.

2. *Data supplying*. The correct use of administrative data depends on the time and conditions of data supplying: delays in the data supplying cause delays in the statistical production and dissemination; changes in legislation "not well disclosed" can determine systematic errors; different moments of data updating among various sources could produce duplications and non-alignment. The solution of these problems are tied to the shape and to the essence of the relationship between statistical and administrative authorities.

3. *The unit delineation.* The most crucial aspect in the problems connected to the implementation of a statistical business register, is that the administrative data - processed and integrated - still refer to legal units. In the UE regulations and EUROSTAT recommendations, the statistical definition of enterprise as "the smallest combination of legal units that is an organisational unit…" and the concept of statistical continuity give a

clear indication of how the statistical universe corresponds to a subset of the legal one and how it results from combinations of legal units complying with the criterion of an enterprise as a unique organisational unit. So the use of administrative registers determine mis-matching with definitions of the UE regulation and problems in business demographics analysis.

**4.** *Large and complex enterprise*. Large enterprises have a complex structure: they carry on more than one economic activity, have several locations, are connected with ancillary units, undergo events of merger and de-merger. The use of administrative sources (not easily controlled) to update a so complex informative structure could cause serious problems: in quality of identifiers, errors of under or over unit sizing, classification errors. Istat experiences suggest that large enterprises should be treated by a skilled staff should using all available information (especially statistical ones) including administrative information..

*References*


Abbate C. Garofalo G. (1998) "*Recent innovation in business register and sample survey on enterprises*", IASS/IAOS Conferences, Aguascalientes, Mexico.

Brackstone G. J. (1987): Statistical Issues of Administrative Data: Issues and Challenges" in "Statistical Uses and Administrative Data" – Statistics Canada

Calzaroni M, Giovannini E., Sorce A. (2000): Il Sistema Informativo Statistico sulle Imprese dell'Istat: problemi e potenzialità, XXX riunione scientifica SIS, Firenze

EUROSTAT (1999), Use of administrative sources for business statistics purposes - Handbook of good practices –

EUROSTAT (1997), proceeding of international workshop: "*Use of administrative sources for statistical purposes*", Luxembourg.

EUROSTAT (1996), "*Recommendation manual business register*".

EUROSTAT (1996), "*Le repertoires statistiques d'enterprises: problems et possibilites",* Actes de la 82° conference des DGINS, Vienna.

Fellegi I. Sunter A.B. (1969) "*A theory for record linkage*" Journal of the American Statistical Association, n.40.

Rivière P. (2000) :"Hybrid data collection: towards a general tool for collecting information from heterogeneous sources" 14th Roundtable on Business Survey Frames – Auckland